Physics Based and Data Driven Anomaly Detection Methods Using Vibration Data for Early Gear Damage Detection in Planetary Gearboxes

Lukas Merkle[1], and Martin Dazer[2]

[1,2] *University of Stuttgart, Institute of Machine Components, Stuttgart, BW, 70569, Germany*
lukas.merkle@ima.uni-stuttgart.de
martin.dazer@ima.uni-stuttgart.de

ABSTRACT

Various methods are known for using vibration sensors to distinguish anomalies such as pitting damage from normal operating conditions in gear drives. These methods are capable of detecting severe damage. However, to take full advantage of prognostics and health management (PHM) strategies, it is necessary to detect damage in a very early damage state. The objective of this work is therefore to analyze, improve and expand the known data evaluation methods in order to achieve the earliest possible detection of gear pitting damage during operation. The research question is: What is the smallest detectable pitting size in planetary gears using vibration data, and which methods are best suited for detection? A high-resolution vibration data set from a single-stage helical planetary gearbox is available for this study. The vibrations were recorded at different speeds, torque levels, and sensor positions. The evaluation methods include entirely physics-based methods in the frequency and time domain, especially a selective analysis of characteristic frequencies and sidebands. These methods are supplemented with data driven approaches such as similarity analysis of frequency spectra. In contrast to physics-based methods, data-driven approaches aim to detect deviations of data sets regardless of their origin. To exclude false positives, these approaches inevitably require the use of multiple measurements without damage as a reference. The application of these methods results in the calculation of 130 condition indicators (CIs). This study provides statistical evidence for the detectability of small damage sizes, particularly using data-driven methods. The key findings of the study are that most of the information about the damage is contained in a comparison of full spectra. A comparison of the full spectral range provides a much clearer picture of damage compared to the analysis of individual characteristic frequencies.

1. INTRODUCTION

As Farrar and Doebling (Farrar & Doebling, 1999) postulated back in 1999, the beginnings of damage detection can be traced back almost indefinitely. They wrote about themselves: "It is the authors' speculation that damage or fault detection, as determined by changes in the dynamic properties or response of systems, has been practiced in a qualitative manner, using acoustic techniques, since modern man has used tools." (Farrar & Doebling, 1999). It is therefore not surprising that there is a comprehensive range of sources available in the field of condition monitoring and damage detection. Common methods for vibration-based gearbox condition monitoring have already been summarized and reviewed in many publications (Lebold, McClintic, Campbell, Byington, & Maynard, 2000; Randall, 2011; Sait & Sharaf-Eldeen, 2011; Vishwakarma, Purohit, Harshlata, & Rajput, 2017; Kundu, Darpe, & Kulkarni, 2021). The basic procedures are part of various standards (ISO International Organization for Standardization, 2004, 2014, 2007, 2020). These publications describe the general methodologies underlying this work.

In 1993, Rytter (Rytter, 1993) defined four categories of methods for vibration-based inspection. The categories were defined by Rytter for civil engineering structures. However, they still provide a useful distinction in the field of vibration-based condition monitoring today:

- "Level 1: The method gives a qualitative indication that damage might be present in the structure. (DETECTION)
- Level 2: The method gives information about the probable location of the damage too. (LOCALIZATION)
- Level 3: The method gives information about the size of the damage. (ASSESSMENT)
- Level 4: The method gives information about the actual safety of the structure given a certain damage state. (CONSEQUENCE)" (Rytter, 1993)

These levels according to Rytter are used in this paper to evaluate the results. It is particularly important to distinguish be-

tween methods suitable to achieve level 1 (detection of a damage) and methods suitable to achieve level 3 (classification of the damage size).

In planetary gearboxes, gear damage usually occurs on the tooth flanks of the gear with the most load cycles. This study therefore focuses on the detection of small pitting damage on the sun gear. With high gear ratios, the sun gear is by far the gear subjected to the highest number of load cycles in the gearbox. For damage detection, various approaches can be found in the literature. These differ in the type of pitting generation, in the applied sensors and in the methods of sensor data evaluation.

An important component in early research on gear condition monitoring is the use and further development of time signal condition indicators for gears (NA4, NA4*, FM4, M6A, M8A, NB4, among others) (Lebold et al., 2000; Zhe, Niaoqing, Fengshou, & Guojun, 2011; Hu, Smith, Randall, & Peng, 2016; Sánchez, Lucero, Vásquez, Cerrada, & Cabrera, 2018; Kundu, Darpe, & Kulkarni, 2019; Sarvestani, Rezaeizadeh, Jomehzadeh, & Bigani, 2020). These vibration condition indicators (CIs) are also used in many studies as comparison values for other evaluation methods (Ozturk, Yesilyurt, & Sabuncu, 2010). However, the significance of these basic time signal indicators is noticeably decreasing in more recent publications. This paper uses basic CIs as a basis for comparison and some of the underlying signal processing methods are also applied here. However, the development of a specific parameter is not pursued. The focus is on general findings regarding the information content of the vibration signal.

In more recent publications, pitting damage is predominantly detected using frequency or order spectrum analyses. For this purpose, characteristic frequencies (usually the gear mesh frequency, GMF) and sidebands are considered (Feng & Zuo, 2012; Korka, Bara, Clavac, & Filip, 2017; Hou, Yang, Li, & Liu, 2022; Li, Xiong, Ma, & Zou, 2023; Yakeu Happi, Kouejou, & Anyika Alugongo, 2023). For measurements with operating conditions that change over time or rapid degradation of the gear, short-time Fourier transform (STFT) (Yakeu Happi et al., 2023), wavelet transform (Ozturk et al., 2010), or Wigner-Ville Distribution (WVD) (Chaari, Fakhfakh, & Haddar, 2006) are also used for combined time and frequency analysis. In some cases, envelope analysis using the Hilbert transform is used to demodulate the signal (Teng, Wang, Zhang, Liu, & Ding, 2014; Kuzio, Zimroz, & Wyłomańska, 2023). These methods are addressed and utilized in this paper.

A main task and challenge in the analyzed research papers was to separate the relevant signal components from unwanted signal components (noise). In complex transmission systems, this task poses a major difficulty due to the large amount of interference with various effects.

The vast majority of studies use single-stage spur gearboxes with straight teeth for their experiments. Helical gears and planetary gears are only found in (Ozturk et al., 2010; Korka et al., 2017; Qu et al., 2019; Sendlbeck, Fromberger, Otto, & Stahl, 2021; Zhu, Maele, Poletto, Baets, & Gryllias, 2024) (helical gears) and (Feng & Zuo, 2012; Li et al., 2023) (planetary gears). In the case of helical teeth, the tooth contact situation is significantly more complex. In planetary gearboxes, several tooth contacts take place simultaneously. This also results in a considerable gap in research.

Another problem is that the test gearboxes and test gears used are often very simplified in design and therefore cannot provide comparable results for real-world applications. Comparisons with near-production gearboxes were only carried out in (Teng et al., 2014; Sarvestani et al., 2020). This study places great emphasis on a near production experimental setup.

There are various approaches to selecting the measurement frequency. In sources that compare acoustic emission (AE) measurements with those using conventional vibration sensors (Sánchez et al., 2018; Grzeszkowski et al., 2019; Kundu, Darpe, & Kulkarni, 2020), no fundamental disadvantages of either technique are apparent. Both measurement methods are similarly well-suited for detecting surface defects. However, it should be noted that in low-frequency ranges, the signal may be masked by unwanted background effects. Consequently, vibration signals typically undergo extensive signal processing and filtering, which is not essential for AE signals above 20 kHz.

In summary, it can be concluded that research on damage detection based on vibration measurement has been conducted for a long time. Many methods and parameters have been developed and applied in numerous studies. The existing research gap is that these methods have only been successfully tested and validated for very large damage sizes (usually pitting sizes $\gg 4\ \%$ or tooth breakage) or in general with large changes in stiffness in the tooth contact. These large damage sizes usually show clearly visible anomalies even in the raw data (time signal). However, this is not sufficient in the context of prognostics and health management (PHM) where the objective is to identify even very minor damage in the system.

This leads to the research question: *What is the smallest detectable damage size of gear damage in planetary gears using vibration data, and which methods are suitable for detection?*

The overall objective is to enable the implementation of novel PHM approaches, as described in (Gretzinger, Kroner, Henß, Dazer, & Bertsche, 2021; Tevetoglu, Henss, Gretzinger, & Bertsche, 2021). PHM can be used to increase the remaining useful life (RUL) of gear drives. However, this requires the very early detection of damage, which is currently not possible with existing state-of-the-art methods.

The explicit aim of this study is not to develop entirely new

Table 1. Test plan. Considered operating conditions and test run numbers v1 to v12.

		Torque T in Nm			
		60	110	160	210
Speed n in min^{-1}	1140	v1	v2	v3	v4
	1670	v5	v6	v7	v8
	2200	v9	v10	v11	v12

performance metrics, but rather to systematically evaluate suitable methods for early damage detection. The goal is to systematically investigate which method is best suited to capture and represent information about emerging damage in the signal. Early detection (which is necessary for PHM) has not been addressed in any literature to date. There is a significant research gap in this area, which this paper aims to close.

2. MATERIALS AND TEST METHODS

Test data according to (FVA Forschungvereinigung Antriebstechnik e.V. & Binanzer, 2026; Merkle, Lukas, Binanzer, Lisa, & Dazer, Martin, 2025) are used to apply the methods. Three planetary gearboxes with a gear ratio $i = 10$ are used as test gearboxes. The test runs are conducted under varying rotational speeds and torque levels to assess the influence of operating conditions on the detection performance. The self-heating of the gearbox and the gearbox oil is monitored during the tests. All test runs are carried out in a temperature range $\Theta = 30\ ^\circ\mathrm{C}...35\ ^\circ\mathrm{C}$. The speed and torque levels of the test runs considered in this study are summarized in Table 1.

For damage detection, three high-frequency acceleration sensors are used, with a sampling rate $f_\mathrm{s} = 250\ \mathrm{kHz}$ and a measurement duration $\Delta t = 100\ \mathrm{s}$ per operating condition. The sensors are mounted at the ring gear, at the bearing of the sun gear, and on the adapter flange of the gearbox. All sensors measure in the radial direction. Based on findings in the literature (see Section 1), the sampling frequency f_s has been set as high as possible to avoid unwanted background effects. The sensors are rated for measuring at least $f_\mathrm{max} = 100\ \mathrm{kHz}$.

In order to be able to compare the recorded vibration data with and without damage, reference measurements without damage are carried out for each gearbox and for each operating condition. For all gearboxes, multiple replications of measurement data (under all operating conditions) are available in undamaged condition. For gearbox 3, there are also two individual sets of measurement data available for each damage level.

Artificial pitting damage is applied to one tooth flank of the sun gear with increasing damage size. The pitting damage is manufactured corresponding to (Häderle, Merkle, & Dazer, 2024) using a numerically controlled milling machine. Besides the undamaged reference state, there are four damage states with damage sizes $A_\mathrm{pit,s} = 0.5\ \%$, $A_\mathrm{pit,m} = 1\ \%$, $A_\mathrm{pit,l} = 2\ \%$ and $A_\mathrm{pit,xl} = 4\ \%$ compared to the area of the tooth flank. Failure criterion according to ISO 6336-5 (ISO International Organization for Standardization, 2016) is $A_\mathrm{pit,fail} = 4\ \%$.

Since the pitting damage is manufactured artificially, the gearbox is disassembled and reassembled between each test series. The gearbox is also disassembled and reassembled between the reference test series without damage in order to be able to identify any effect the process may have on the data.

3. EVALUATION METHOD

The evaluation of the measurement data is structured into the following parts: Preprocessing, time domain signal processing, frequency domain signal processing and the calculation of 1-dimensional and N-dimensional CIs.

The term "CI" now summarizes various methods. These include, on the one hand, the traditional basic CIs (based on the time signal) and, on the other hand, newly developed approaches. These can be divided into methods directly based on physical relationships and methods based purely on a comparison of available data. The distinction between physics-based and data-driven methods is not perfectly clear-cut. In general, only data processing steps are used that are generally appropriate in the context of vibration measurements. The determination of characteristic values at characteristic gear frequencies, for example, is associated with physics-based methods, while the calculation of a difference spectrum is more commonly associated with data-driven methods. CIs characterize the health state (condition) of the transmission and can consist of a single value (1-dimensional CI) or multiple values (N-dimensional CI). Thus, entire signals or spectra can also be considered CIs. If 1-dimensional CIs are calculated from the N-dimensional CIs, these are referred to as "resulting CIs".

3.1. Signal Preprocessing

The acquisition of vibration data and incremental encoder data is shown in Figure 1 (illustration based on (Merkle, Lukas et al., 2025)). The vibration sensor is a piezoelectric sensor that outputs accelerations in the form of a voltage-time signal. The incremental encoder counts 10 000 steps per revolution as well as an index signal at one position per revolution. During disassembly and reassembly of the gearbox, care was taken to align the index signals of both encoders with the exact position of the gearbox input and output shafts. The incremental encoder provides the index signal once per full revolution of the shaft as reference position. The index signal is also used for calculating the mean rotational speed in the measurement interval.

The initial signal preprocessing is conducted according to the state of the art (Lebold et al., 2000), see Figure 2. The first

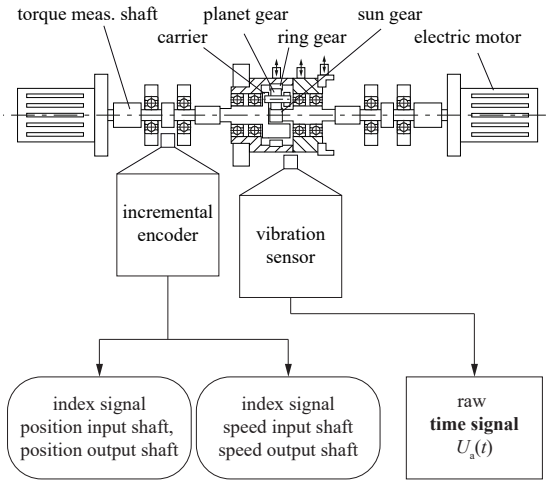

Figure 1. Data acquisition on the test bench.

step is signal conditioning of the raw voltage signal. The signal is multiplied by a constant factor (here, the sensor sensitivity S) and the constant DC (direct current) bias is removed. The result is a conditioned raw signal. The signal is now available as an acceleration-time signal.

Downstream, filtering (denoising) and/or windowing in the time and/or frequency domain could be possible. This type of filter is avoided in this step. Indirect filtering by removing stochastic signal components using time-synchronous averaging (TSA) (see Section 3.2) is used in further steps. Random signal components are removed, while recurring components remain in the signal (Bechhoefer & Kingsley, 2009).

In addition to vibration and rotation angle information, temperatures and torques are also recorded during the test runs to verify the operating conditions. For damage detection, only the vibration measurements will be considered further in this paper.

3.2. Time Domain Signal Processing

Basic signal processing is carried out in the time domain by demodulation and TSA. This leads to three additional time signals derived from the preprocessed time signal, see Figure 3. TSA divides the time signal into physically meaningful sections (here, one revolution of the carrier or 10 revolutions of the sun gear). Because the gear ratio is an integer ($i = 10$), the gearbox returns to its initial position after one revolution of the carrier and 10 revolutions of the sun gear. Averaging attenuates random signal components, resulting in a short signal with mainly the periodically recurring components (Bechhoefer & Kingsley, 2009). The index signals from the incremental encoders are used to calculate the TSA signal. The envelope (calculated here by the absolute value of the Hilbert transform) demodulates the (amplitude-modulated) signal.

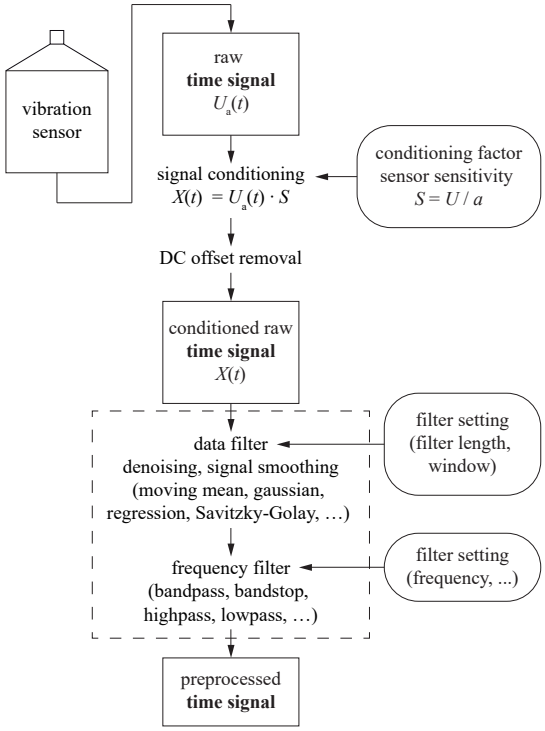

Figure 2. Signal preprocessing.

The downstream windowing of the signals is designed globally to be suitable for all measurements in order to enable comparison. Matching signal lengths generate equal frequency bins. Identical signal data point lengths could also be generated by resampling rather than windowing. However, this can result in interpolation errors or aliasing effects. The window length of the applied windowing function is approximately $99.9\ \%$ of the signal length. This results in a smaller change to the signal than resampling. Calculating the (upper) envelope creates a DC offset again, which is removed for all signals used to calculate spectra.

A TSA signal extended by $10\ \%$ on both sides (total TSA signal length $120\ \%$) is generated, in order to be able to compensate for the repeatability error of the installation process. The beginning and end areas overlap. This signal is necessary in order to have enough signal length for a later performed phase shift for analysis in the time domain. The gearbox must be disassembled and reassembled between pitting sizes. In this $120\ \%$ signal the DC bias is not removed.

3.3. 1-Dimensional Time Domain Signal Condition Indicators

The simplest form of damage detection is performed by calculating 1-dimensional parameters directly with the time signal, see Figure 4. Typically, the parameters are calculated from the TSA signal or the raw signal (or both). Here, the calculation is conducted with all time signals.

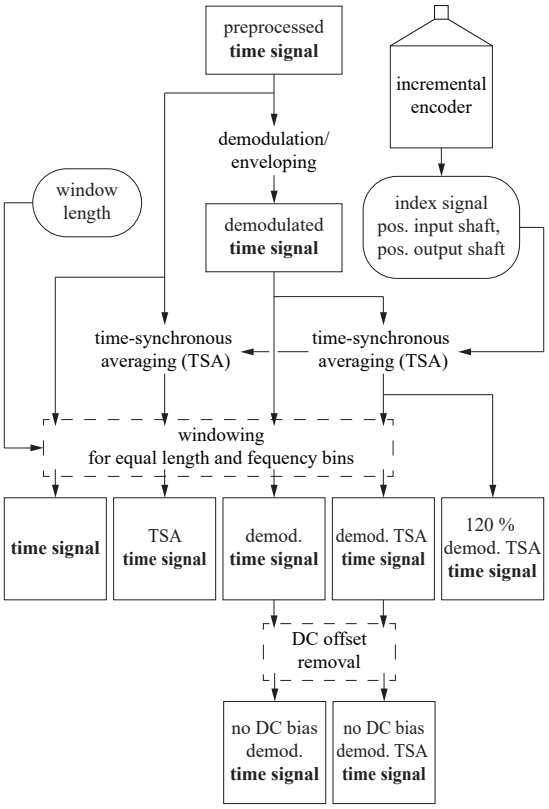

Figure 3. Time domain signal processing.

In the time domain, the calculation of a "residual signal" or "difference signal" is omitted, which means that some of the traditional (Lebold et al., 2000) CIs cannot be calculated at this stage. For example, NA4, NA4* FM4, M6A, M8A, and NB4 are not used. As described in Section 1, the use of these CIs in modern research is declining considerably. However, root mean square (RMS), kurtosis, crest factor (CF), and the peak-to-peak values are calculated for general classification and comparison purposes.

3.4. Frequency Domain Signal Processing

The fast Fourier transform (FFT) transfers the time domain signal into the frequency domain. Generally, every time signal can be used for this purpose. In addition to the one-sided amplitude spectrum, power and cepstrum can also be calculated using the FFT, see Figure 5.

The power spectrum is calculated directly from the FFT by scaling (valid for a rectangular window and discrete power spectra). The power spectral density (PSD) is determined by dividing the power spectrum by the equivalent noise bandwidth (ENBW). For rectangular windows with a window length L, the ENBW is $ENBW = f_\mathrm{s}/L$. The second method is to use Welch's method (Welch, 1967) as an approach for PSD estimation. This achieves noise reduction while simultaneously reducing frequency resolution. The following parameters are used for Welch's PSD estimation:

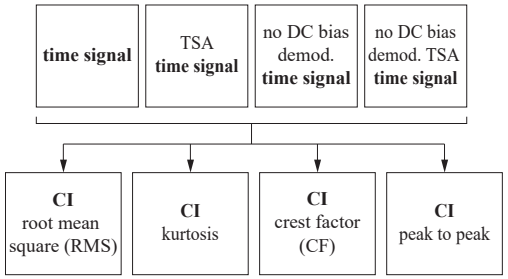

Figure 4. 1-dimensional time domain signal CIs.

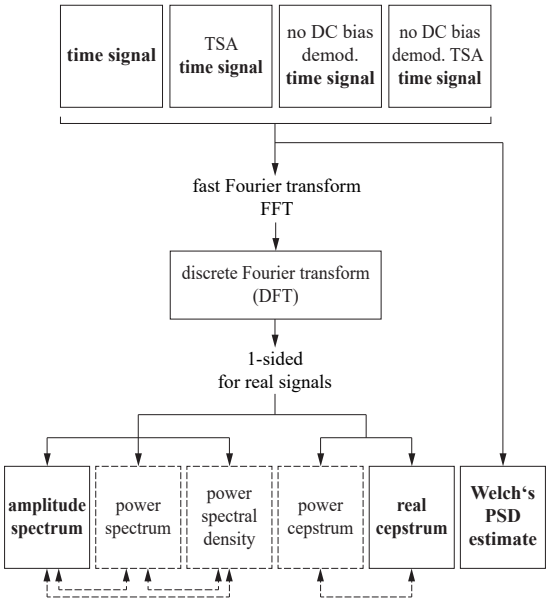

Figure 5. Frequency domain signal processing.

1. The window type is set to Hamming window (default setting in (The MathWorks, Inc., 2026))
   $ENBW_{\mathrm{Hamming}} = 1.36...1.37$
2. The standard window length is $L_\mathrm{W} = L/4.5$ to get the longest possible segments out of the signal length L to obtain as close to but not exceed 8 segments with $50\ \%$ overlap (default setting in (The MathWorks, Inc., 2026))
3. If the signal is long enough the window length is defined such that the resulting resolution bandwidth (RBW) is $50\ \%$ of the lowest frequency of interest (rotation frequency of carrier) $RBW = 0.5 \cdot f_{\mathrm{carrier}}$. The window length is therefore $L_\mathrm{W} = ENBW \cdot f_\mathrm{s}/RBW = ENBW \cdot f_\mathrm{s}/(0.5 \cdot f_{\mathrm{carrier}})$
4. The window overlap is set to $50\ \%$
5. The number of discrete Fourier transform (DFT) points to use in the PSD estimate is set to the full signal length L to achieve equal frequency bins compared to the single sided amplitude spectrum

The window length (2., 3.) influences the frequency RBW. The RBW indicates which frequency widths in the spectrum are still distinguishable. In contrast to this, increasing the number of DFT points to L (5.) has no effect on the information content of the PSD estimation, but is used for better comparability of the results with equal frequency bins.

The cepstrum C_p (Bogert, Healy, & Tukey, 1963) can also be defined as the "spectrum of the spectrum" (Norton, 2003), see Eq. 1. $\mathcal{F}$ is the Fourier transform and $X(t)$ is the vibration signal.

$$C_\mathrm{p} = \left| \mathcal{F} \left\{ \log \left(|\mathcal{F}\{X(t)\}|^2 \right) \right\} \right|^2 \tag{1}$$

Therefore, as with the spectrum of a time signal, there is an effect equivalent to DC offset, visible in the spectrum at frequency $f = 0$ or in the cepstrum at quefrency $q = 0$. In time signals, this DC effect is removed. In the cepstrum, the offset is also removed accordingly.

Comparably to the power spectrum, the power cepstrum can be derived directly from the real cepstrum. Although the power spectrum, PSD, and power cepstrum are calculated, they do not provide any additional information beyond the amplitude spectrum and the real cepstrum. They are excluded from further evaluation and are therefore enclosed in a dashed box in the graph in Figure 5. Since only constant rotational speeds are considered in this study and a comparison is always made within the same operating conditions, the evaluation is limited to the frequency range. An order spectrum analysis is not performed.

3.5. Search Method for Frequencies in the Spectrum

All periodic events in the gearbox that are relevant for detecting tooth flank damage depend on the number of teeth z, the number of planets q_planet and on the rotational frequencies. They can therefore be expressed as a multiple of the input or output rotational frequency f_carrier or f_sun.

The GMF is calculated according to Eq. 2 or Eq. 3.

$$f_\mathrm{GMF} = f_\mathrm{carrier} \cdot z_\mathrm{ring} \tag{2}$$

$$f_\mathrm{GMF} = f_\mathrm{sun} \cdot \frac{z_\mathrm{ring} \cdot z_\mathrm{sun}}{z_\mathrm{ring} + z_\mathrm{sun}} \tag{3}$$

The frequency of the damaged tooth in contact is calculated according to Eq. 4 or Eq. 5.

$$f_\mathrm{damage} = f_\mathrm{carrier} \cdot q_\mathrm{planet} \cdot \frac{z_\mathrm{ring}}{z_\mathrm{sun}} \tag{4}$$

$$f_\mathrm{damage} = f_\mathrm{sun} \cdot q_\mathrm{planet} \cdot \frac{z_\mathrm{ring}}{z_\mathrm{ring} + z_\mathrm{sun}} \tag{5}$$

The frequency of the bypassing planets is calculated according to Eq. 6 or Eq. 7.

$$f_\mathrm{pass\ planet} = f_\mathrm{carrier} \cdot q_\mathrm{planet} \tag{6}$$

$$f_\mathrm{pass\ planet} = f_\mathrm{sun} \cdot q_\mathrm{planet} \cdot \frac{z_\mathrm{sun}}{z_\mathrm{ring} + z_\mathrm{sun}} \tag{7}$$

To extract individual values from a spectrum, a new procedure is developed in this work. Due to imperfections in the rotational speed, a finite frequency resolution and finite frequency bins, the amplitude values at the corresponding frequencies cannot simply be read off. Instead, the peaks in the spectrum must be "searched" for. The problem intensifies very quickly if harmonic multiples are to be reliably found in the high frequency ranges as well.

To determine amplitude values, a search window is opened around the calculated frequencies. The window width is $\pm 10\ \%$ of the lowest frequency of interest (rotation frequency of carrier f_carrier). The window width therefore depends on the rotational speed of the gearbox.

The maximum amplitude is searched for within this window width and stored in a table together with the actual frequency value at the point of maximum. Due to the discrete sampling and minimal discontinuities in the operating point, the calculated frequencies may differ slightly from the frequencies actually found. To prevent the calculated and actual frequencies from increasingly diverging in the higher frequency ranges, the calculated frequencies are adjusted based on the first $1\ \%$ of the data using a linear fit. The calculated and the actual frequencies are compared with each other, and the correlation is used as a correction value for further frequency calculation.

In terms of principle, there is a maximum in every search window. However, due to the multiple superimposition of effects in the signal, an incorrect frequency may also be found. Therefore, the position of each value is automatically double checked. If the frequency at the found peak f deviates from the calculated frequency f_cal by more than $\pm 50\ \%$ of the search window width, the found characteristic value is rejected and set to NaN (not a number).

3.6. 1-Dimensional Frequency Domain Condition Indicators

Typical frequencies can be determined for every gearbox and every operating condition. These frequencies result directly from the gear kinematics and are therefore unique and can be clearly assigned to the rotation and tooth contacts in the gearbox. These frequencies are therefore referred to as "characteristic frequencies" in the following. The corresponding values are "characteristic values". The same applies to the quefrencies in the cepstrum. The characteristic quefrency can

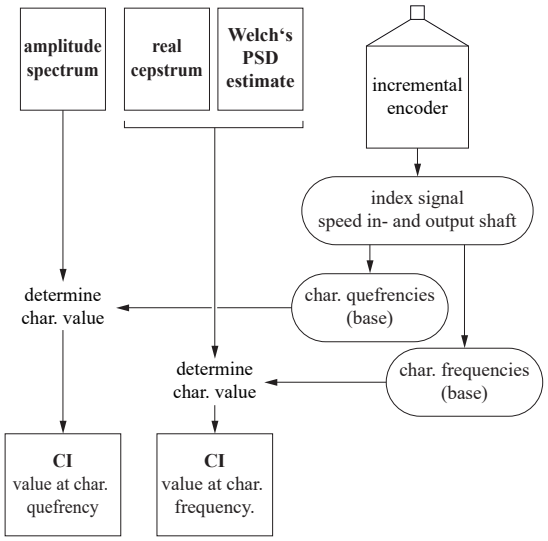

Figure 6. 1-dimensional frequency based CIs.

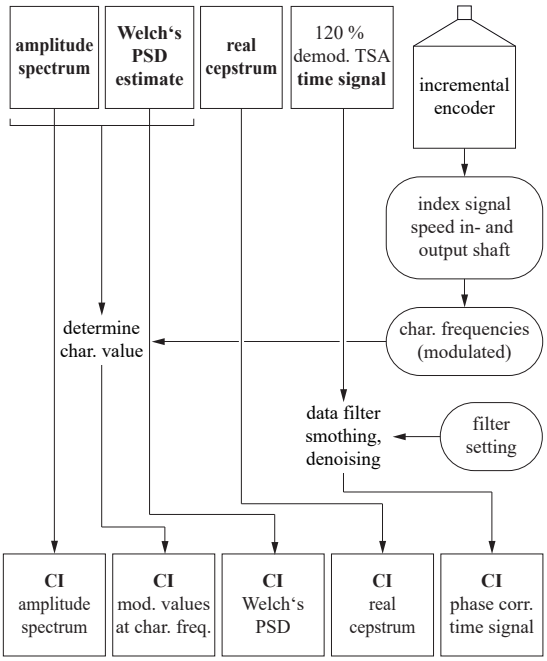

Figure 7. N-dimensional frequency domain based CIs.

be calculated as the reciprocal of the corresponding characteristic frequency and is given in seconds.

These frequencies are calculated using the incremental encoder signals and the gear data and searched for in the frequency spectrum (see Section 3.5). These frequencies can be used directly as CIs, see Figure 6. Of particular interest are the GMF f_{GMF} and the frequency of the contact of damaged tooth f_{damage}. Due to their low frequency resolution, the spectra of the TSA signals are not suitable for determining characteristic frequencies. For the TSA spectra, the distances between the frequencies searched for are in the same range as the frequency bins.

3.7. N-dimensional Frequency Domain Based Condition Indicators

To compare the damage conditions, N-dimensional CIs are derived from spectra and cepstra and the demodulated TSA time signal is used directly as a characteristic value (with applied phase shift and filtering), see Figure 7. The measured rotational frequencies f_{sun} and f_{carrier} and the kinematic relationships are used to calculate the modulated multiple characteristic frequencies. The values (amplitudes and PSD) at these characteristic frequencies are searched for (see Section 3.5) and stored as characteristic values. If no dominant peak is found, the value is not stored (NaN).

The GMF is a carrier frequency for the low-frequency rotational, planetary, and damage frequencies. These frequencies form upper and lower sidebands around the GMF (Inalpolat & Kahraman, 2009). To speed up the calculation, the effect that the sideband frequencies are integer multiples of each other is also exploited here, and all sidebands fall on the sidebands of the rotational frequency of the carrier, which are evenly distributed across the frequency spectrum.

As in Section 3.6, the following also applies here: Due to their low frequency resolution, the spectra of TSA signals are no longer suitable for determining characteristic frequencies. The distances between the frequencies searched for are in the same range as the frequency bin. The cepstrum as the "spectrum of the spectrum" no longer shows any distinct multiples of the quefrencies and is therefore not used for calculating N-dimensional CIs.

The $120\ \%$ extended TSA time signal is corrected in its phase position for using it as N-dimensional CI using a reference signal. A moving mean filter with a filter length of $10\ \%$ of the signal length is used to determine the phase positions. This has the effect of a low-pass filter. For phase position correction, only dominant low frequency signal components are to be used in order to achieve a more robust result of the phase shift. The phase position is determined by calculating the cross-covariance between the signals. The phase position is applied uniformly and globally. This means, first a signal lag is determined for all measurement data individually. Then, the lags of all measurements between which no disassembly and reassembly took place are summarized and the median is applied to all corresponding measurements. The median is not influenced by individual outliners. If no disassembly and reassembly has taken place, there can be no deviation in the phase positions. After the phase shift, the signal is shortened again to the length one revolution of the carrier.

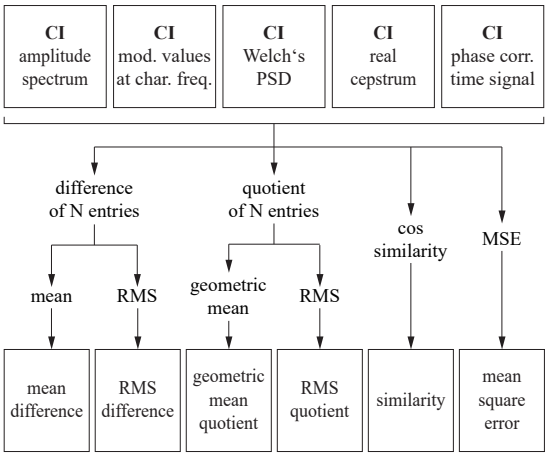

Figure 8. Calculation of resulting 1-dimensional CIs with N-dimensional CIs.

3.8. Resulting 1-Dimensional Condition Indicators Calculation with N-Dimensional Condition Indicators

For a quick overview of health status, the N-dimensional CIs are reduced to resulting 1-dimensional CIs, see Figure 8.

A special feature of the procedure developed here is the direct comparison of measurements with damage and the corresponding reference measurements without damage. A major difference from the state of the art is that a condition indicator is not simply calculated for each measurement, but rather the N-dimensional CIs are calculated "line by line". For example, the amplitude of a measurement with damage at the frequency $f_\mathrm{a} = 300$ Hz is compared with the amplitude at $f_\mathrm{b} = 300$ Hz in the healthy measurement. The calculation of the resulting CI therefore always requires a reference basis. The reference basis can also be an average of several measurements. If a rejected value (NaN) is obtained in one or more entries of a row, the entire row is discarded. A valid value for each frequency must be found for all measurements. Only then is a comparison based on this value possible. Using this method, only very few "random" values in the center of the search remain in the measurement data.

To determine the resulting CIs, the signed change is calculated by forming differences. In this case, $\vec{a}$ is a condition with damage and $\vec{b}$ is the corresponding reference (Eq. 8).

$$\overrightarrow{CI_\mathrm{res}} = \begin{pmatrix} a_1 - b_1 \\ a_2 - b_2 \\ \vdots \\ a_n - b_n \end{pmatrix} \tag{8}$$

The relative change is calculated accordingly by forming element-wise quotients (Eq. 9).

$$\overrightarrow{CI_\mathrm{res}} = \begin{pmatrix} a_1/b_1 \\ a_2/b_2 \\ \vdots \\ a_n/b_n \end{pmatrix} \tag{9}$$

To reduce an N-dimensional CI to a 1-dimensional resulting CI, for the signed difference the mean is calculated and for the relative change (quotient) the geometric mean is calculated. The RMS is also calculated for both values, Eq. 10. The RMS is independent of the sign and more sensitive to larger values.

$$RMS = \sqrt{\frac{1}{n}\sum_{i=1}^{n} CI_{\mathrm{res},i}^2} \tag{10}$$

To calculate the geometric mean, the values must be positive. For cepstral and spectral values, this is ensured by the definition used here. The phase-shifted TSA time signal is a demodulated signal. The DC-bias is not removed afterwards, so this is also a signal with exclusively positive values.

For a straightforward similarity analysis of the CIs, the cosine similarity is calculated (Eq. 11).

$$cosSim = \cos(\theta) = \frac{a \cdot b}{\|a\|\,\|b\|} = \frac{a \cdot b}{\sqrt{a^2 \cdot b^2}} \tag{11}$$

The mean squared error (MSE) is also used to calculate a resulting CI (Eq. 12).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \text{ ... here: } \frac{1}{n}\sum_{i=1}^{n}(a_i - b_i)^2 \tag{12}$$

Calculating the RMS of the signed difference results in calculating the Root Mean Squared Error (RMSE) (Eq. 13).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_i - b_i)^2} \tag{13}$$

All measurements of an operational condition are compared against a common reference base. The reference base is the average of all healthy measurements. In order to be able to compare each measurement in healthy state against the reference base, each measurement in healthy state is sequentially excluded from the reference base to obtain an independent reference or comparison.

3.9. Assessment Scale for the Suitability of Condition Indicators

In order to be able to quickly evaluate the large number of 130 1-dimensional CIs and resulting CIs, two evaluation criteria are defined. These are based on levels 1 and 3 according to Rytter (Rytter, 1993).

- Level 1 (DETECTION): qualitative indication that **damage might be present**.
- Level 3 (ASSESSMENT): information about the **size of the damage**

To quantify the detection capability of a CI, a statistical t-test is carried out. The two test groups are group 1 (CI values of all damaged, y_{damaged}) and group 2 (CI values of all undamaged, y_{healthy}). All CIs with p-values $p > 0.05$ indicate no statistically significant change in the value. They are sorted out. All remaining CIs are evaluated based on their effect size d (Eq. 14).

$$d = \frac{|\mathrm{mean}(y_{\mathrm{damaged}}) - \mathrm{mean}(y_{\mathrm{healthy}})|}{\sqrt{\frac{\mathrm{std}(y_{\mathrm{healthy}})^2 + \mathrm{std}(y_{\mathrm{damaged}})^2}{2}}} \tag{14}$$

To quantify the ability of a parameter to classify the damage size (assessment), a linear model is fitted to the CI and the size. As above, all CIs with p-values $p > 0.05$ are not significant and sorted out in the first step. The model accuracy, and with this the ability to classify the damage size is measured using the R^2 value.

Each group includes all tests, all test gearboxes, and all sensor positions. The objective is to identify the overall best CI for all conditions.

3.10. Combined Time and Frequency Approaches

Due to the constant operating conditions during the measurement period, combined methods in the time and frequency domain are of secondary importance in damage detection in this work. Instead of the combined analysis, it is better to use the full signal length for spectrum formation, which results in very narrow frequency bins. The time signal is only divided when forming the TSA signal for reasons of principle.

However, if, for unknown reasons, the signal were nevertheless transient, this would have a significant negative impact on the evaluation. Since it is not possible to assess whether a signal is transient based on the spectra, each signal is displayed as a persistence spectrum for visual inspection and checked for plausibility.

4. RESULTS

The results of the intermediate steps of the evaluation are shown here in detail for a single measurement only as an example. For the actual damage detection, only the corresponding best CIs are displayed.

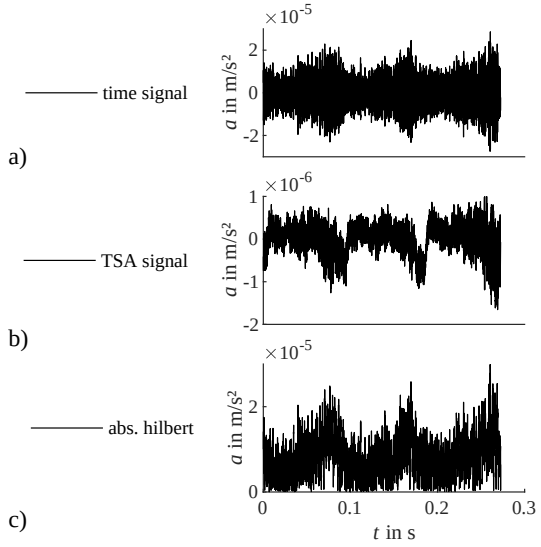

Figure 9. a) Time signal, shortened to one rotation of the carrier (10 rotations of the sun gear). b) TSA signal (enlarged). c) Upper envelope signal. Test run v12, gearbox 2.

4.1. Signal Processing

To illustrate the presented damage detection approach, the test run v12 $n = 2,200\ \mathrm{min}^{-1}$, $T = 210\ \mathrm{Nm}$ is examined in more detail below. The vibration measurements of the sensor on the ring gear are shown as an example.

The characteristic gear frequencies measured in this test are as follows:

- $f_{\mathrm{sun}} = 36.6474\ \mathrm{Hz}$
- $f_{\mathrm{carrier}} = 3.6647\ \mathrm{Hz}$
- $f_{\mathrm{GMF}} = 395.7918\ \mathrm{Hz}$
- $f_{\mathrm{pass\ planet}} = 10.9942\ \mathrm{Hz}$
- $f_{\mathrm{damage}} = 98.9479\ \mathrm{Hz}$

Figure 9 shows the waveform of the measured time signal (conditioned raw signal). The 100-second measurement duration has been reduced to the duration of one revolution of the carrier (10 revolutions of the sun gear). The passage of the three planetary gears in the gearbox is clearly visible, particularly in the measurement data from the sensor on the ring gear. The passing planets are also clearly visible in the TSA signal and in the (demodulated) envelope signal.

The characteristic frequencies (base) of the gearbox are visible in the lower part of the frequency spectrum, see Figure 10. In the enlarged view, it is clear that all frequencies have been calculated correctly and they are clearly visible. However, in the overall spectrum, these frequencies are generally overshadowed in amplitude by a few neighboring, more dominant frequencies. To account for these effects, the window width

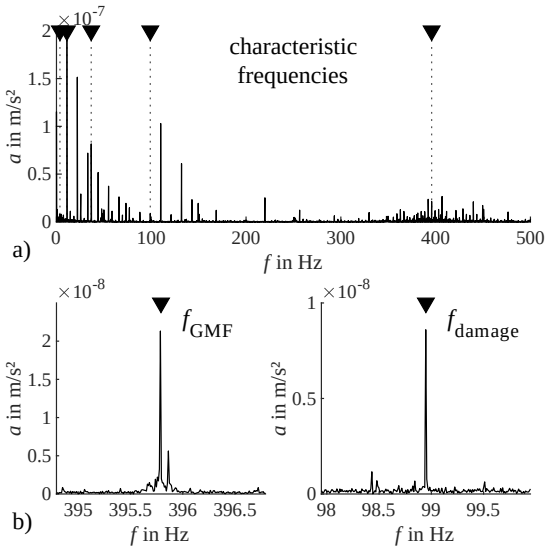

Figure 10. Characteristic base frequencies as 1-dimensional CI. a) All frequencies. b) The frequencies relevant in this paper f_{GMF} and f_{damage} enlarged. Test run v12, gearbox 2.

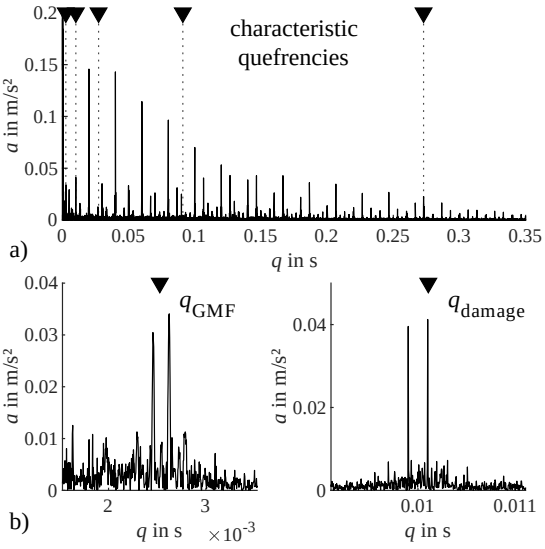

Figure 12. Characteristic quefrencies as 1-dimensional CI. a) All quefrencies. b) The quefrencies relevant in this paper q_{GMF} and q_{damage} enlarged. Test run v12, gearbox 2.

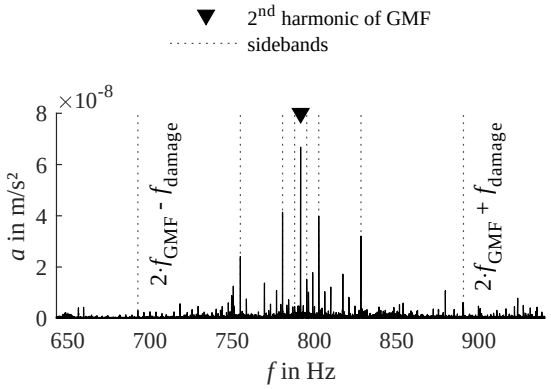

Figure 11. Example of modulated frequencies in the amplitude spectrum around the 2nd harmonic of the GMF. Test run v12, gearbox 2.

was selected as described in Section 3.5.

The modulated frequencies are illustrated using the example of the second harmonic of the GMF, see Figure 11. To the right and left of this harmonic, the upper and lower sidebands appear at the same intervals as the base frequencies. Shown here is the first of several sidebands for each frequency. It should also be noted, however, that some sidebands overlap at the same frequency, since these frequencies are integer multiples of one another.

Characteristic values can also be determined in the cepstrum. In contrast to the spectrum, the values in the cepstrum are transformed back to the time domain ("quefrency"). These are the reciprocals of the characteristic frequencies. In contrast to the frequency spectrum, the characteristic values in the cepstrum are much less pronounced and are significantly overshadowed by other values. The search algorithm with a search window is applied in the cepstrum in the same way as in the frequency spectrum.

The PSD calculated using Welch's method and the parameters mentioned in Section 3.4 results in an averaging of the power within the RBW range, see Figure 13. The sidebands can still be distinguished without difficulty. However, the power spectrum is heavily filtered.

4.2. Search Windows and Characteristic Frequencies

To illustrate the search windows used to calculate the characteristic values, the second harmonic of the GMF is plotted with the sidebands of the planet pass frequency, see Figure 14, top. The applied search windows with a width of ± 10 % of the lowest frequency f_{carrier} are plotted around the calculated frequencies. The maximum value is determined within this search window. This principle is applied to all harmonics across the entire frequency spectrum. Characteristic values are calculated across the entire spectrum. No restrictions are initially applied.

To verify the position of the frequencies, the difference between the calculated frequency and the actual frequency is calculated and plotted against the calculated frequency, see Figure 14, bottom. In the low-frequency range up to 10 kHz, the calculated frequencies and the actual frequencies are very close to each other. The maximum is found predominantly in the center of the search window. All values that deviate by more than 50 % from the center of the search window are rejected, as described in Section 3.5.

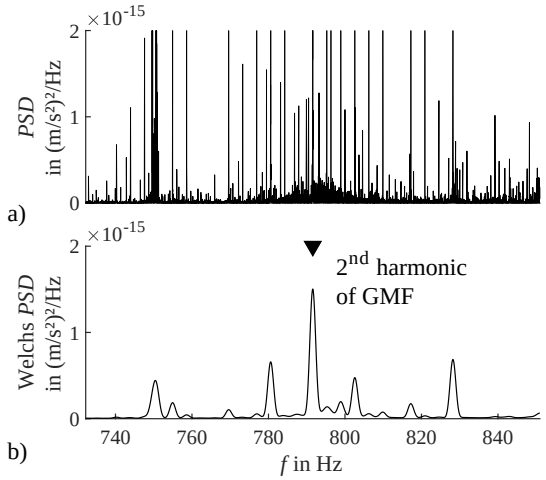

Figure 13. a) PSD calculated with the FFT. b) PSD estimate using Welch's method. Values around the 2nd harmonic of the GMF. Test run v12, gearbox 2.

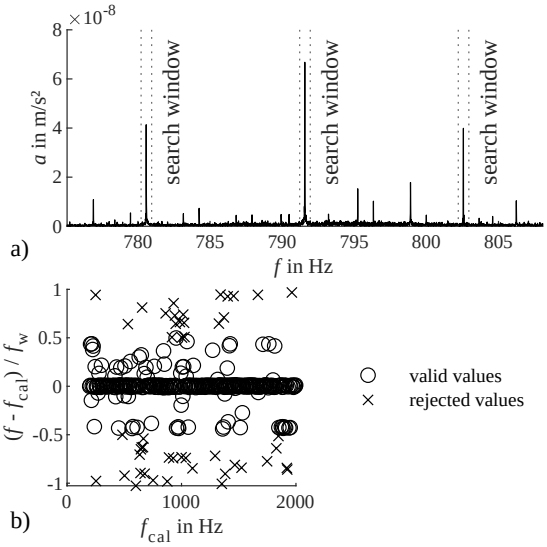

Figure 14. a) Search window for characteristic frequencies. b) Accepted and rejected values depending on the position of the maximum in the window. Both: Exemplary frequency ranges. Test run v12, gearbox 2.

A look at the entire frequency spectrum quickly reveals that, up to approximately 20 kHz, there is a trend away from frequencies concentrated in the center toward a random distribution. It can therefore no longer be assumed that actual peaks exist at the calculated frequencies.

By combining all measurements under a single operating condition, all frequencies with rejected values are filtered out. This reduces the full spectrum $f_{\max} = 125$ kHz to characteristic values found primarily in the range up to $f_{\max,\mathrm{char}} = 20$ kHz, see Figure 15.

4.3. Persistence Spectrum for Signal Stationarity Check

Figure 16 shows an example of the persistence spectrum of the time signal and the TSA time signal. The power is given in dB. A sharp line is visible in both the TSA signal and the time signal. It can therefore be assumed that the operating conditions are completely stationary. The persistence spectra are verified for all measurements.

4.4. Evaluation of Condition Indicators

The objective of this research work is to identify an evaluation method that not only performs well for a specific gearbox or a specific sensor or within specific operating conditions, but also delivers good results across all measurements.

The results of the t-test show 15 CIs with significant p-values and $d > 0.5$ and 5 CIs with significant p-values and $d > 0.8$. According to (Cohen, 1988), the effect size can be considered "medium" if $d > 0.5$ and "large" if $d > 0.8$, see Figure 17 and Table 3 (Appendix). It is therefore statistically possible to detect the damage.

The five CIs with the greatest effect size are (numbering corresponding to the Figure 18):

a) The geometric mean quotient of the CI amplitude spectrum of the demodulated TSA time signal without DC bias (DM TSA TS AS gmean div)
b) The mean difference of the CI amplitude spectrum of the demodulated time signal without DC bias (DM TS AS mean dif)
c) The geometric mean quotient of the CI amplitude spectrum of the demodulated time signal without DC bias (DM TS AS gmean div)
d) The mean difference of the CI amplitude spectrum of the time signal (TS AS mean dif)
e) The geometric mean quotient of the CI amplitude spectrum of the time signal (TS AS gmean div)

Figure 18 shows the distribution of the resulting CIs in detail. For each of the three gearboxes, the resulting CIs are plotted for all sensors above the damage size. The histogram on the right shows the two groups of the t-test: "healthy" and "damaged". Since multiple measurements are available for each gearbox in an undamaged state (the mean value of the independent measurements is used as the reference), all undamaged measurements can be plotted at a damage size of 0. Damage detection is possible if all data points with damage are clearly different from the measurements without damage.

With $d = 0.969$, the resulting CI DM TSA TS AS gmean div has the largest effect size. The comparison with the resulting CI DM TS AS gmean div with $d = 0.944$ is notable. By using a TSA signal, a measurably larger effect size can be achieved in this case. A further comparison with the resulting CI TS AS gmean div shows that the information content can also be significantly increased through demodulation. The envelope

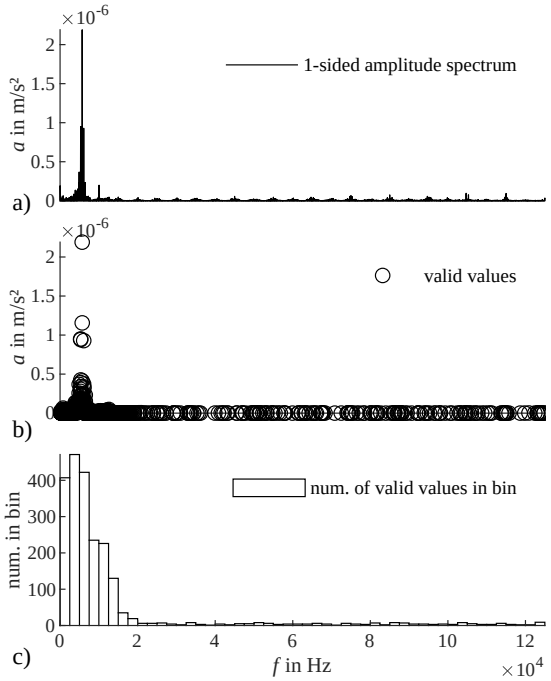

Figure 15. a) Full spectrum (one-sided amplitude spectrum). b) Remaining valid characteristic values. c) Histogram of the remaining valid values per frequency bin. Test run v12, gearbox 2.

spectrum has a 13 % greater effect size compared to the CI based on the amplitude spectrum. In contrast, the comparison between DM TS AS mean dif and DM TS AS gmean div shows that examining the ratio of frequencies gives approximately the same results as examining the differences. The extensive calculation of the characteristic frequencies for the resulting CI DM TS AS CFm mean dif offers no advantage here over the analysis of full spectra.

Even in the best-case scenario (for the geometric mean of the CI amplitude spectrum of the demodulated TSA time signal without DC bias), the linear model only achieves $R^2_{\max} = 0.0677$. This means classification of different damage sizes is not possible.

4.5. Comparison with Basic Condition Indicators According to the State of the Art

To help interpret the results shown, the three common metrics (RMS of the time signal, kurtosis of the time signal, and the amplitude value of the GMF) are also presented in Figure 19. It is immediately apparent that there is no recognizable correlation between the values and the extent or presence of the damage.

See Table 2 for the results of the t-test. For TS kurtosis and AS TS RMS, the p-values are > 0.05 and are therefore not significant. For f_{GMF}, although the p-value is small, the effect size of $d = 0.220$ is low and therefore also not useful.

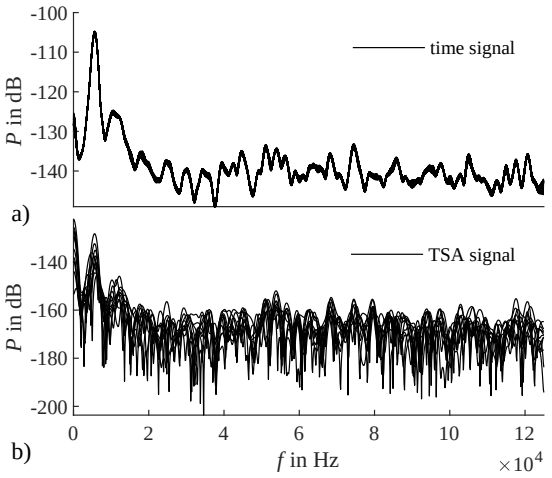

Figure 16. Persistence spectrum. a) Of the time signal. b) of the TSA signal. Test run v12, gearbox 2.

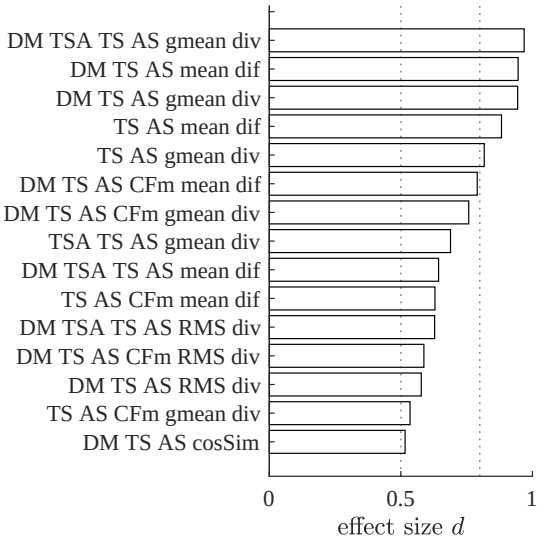

Figure 17. CIs with significant p-values and $d > 0.5$.

4.6. N-Dimensional Condition Indicator

To illustrate the behavior of N-dimensional CIs, the best resulting CI DM TSA TS AS gmean div is shown in detail, Figure 20. Again, the evaluation of the vibration measurements of the sensor on the ring gear of test run v12 $n = 2,200\ \mathrm{min}^{-1}$, $T = 210\ \mathrm{Nm}$ is shown as an example. The graph shows the quotients relative to the reference value for measurements without damage and measurements with damage size $A_{\mathrm{pit,xl}} = 4\ \%$.

When examining the relative change of the N-dimensional CI DM TSA TS AS div, at first sight, the graph looks very similar. To illustrate the trend of the geometric mean across the frequency range, it is shown as a moving geometric mean. In the range up to 30 kHz, the values from the measurement

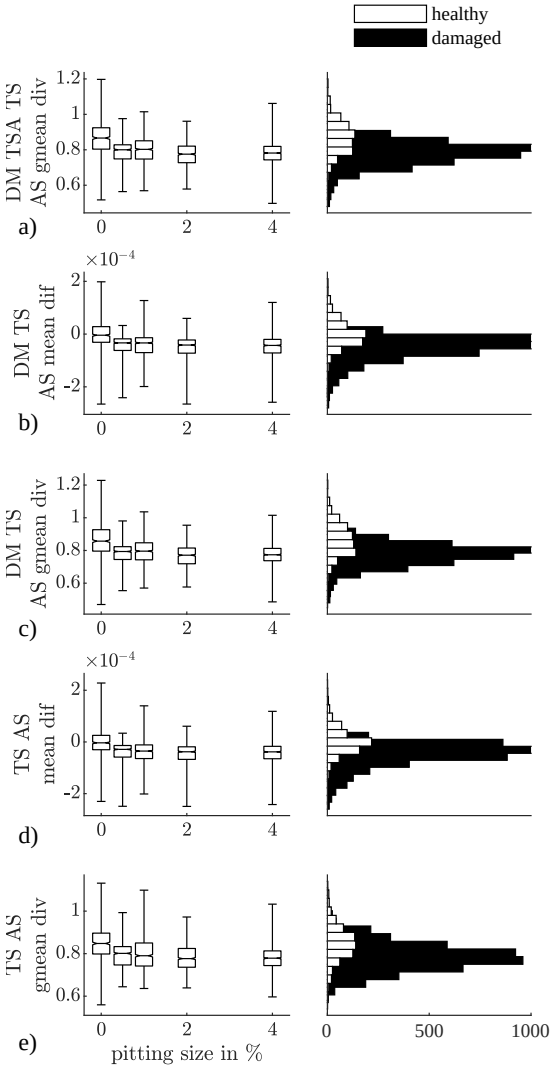

Figure 18. Resulting CI over damage size. a) - e) Five best performing CIs, all three sensors and all three gearboxes.

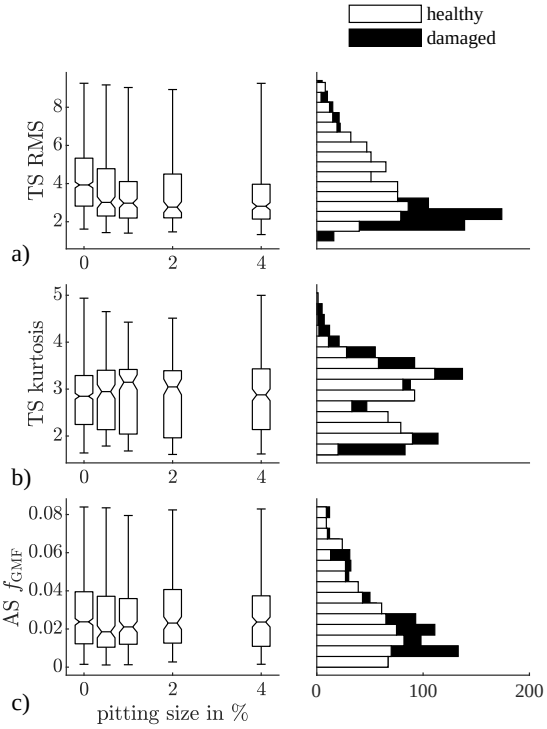

Figure 19. CI over damage size for comparison. a) Time signal RMS. b) Kurtosis. c) The GMF in the amplitude spectrum.

Table 2. p-value and effect size d of the three common metrics RMS of the time signal, kurtosis of the time signal, and the amplitude value of the GMF

CI	p-value	effect size d
TS RMS	0.847	0.010
TS kurtosis	0.381	0.045
AS f_{GMF}	< 0.001	0.220

without damage are significantly higher than those from the measurement with damage. Later on, the curves approximate each other. However, the curve from the measurement without damage is almost always above that of the curve with damage. This is ultimately reflected in the resulting 1-dim CIs (horizontal lines in Figure 20). The fact that low-frequency ranges behave differently from high-frequency ranges is consistent with findings in literature (Kundu et al., 2020; Grzeszkowski et al., 2019; Sánchez et al., 2018). They used high sampling frequencies for avoiding undesirable background effects and obtaining stable measurement values.

5. DISCUSSION

The most important finding of this study is that it is possible to identify CIs that enable early damage detection in planetary gearboxes. Upon closer examination, it becomes apparent that detection is only possible through a comparison of full spectra with corresponding spectra from a healthy state. The characteristic frequencies of the gearbox (base frequencies and multiples modulated frequencies with sidebands) often cited in the literature show less noticeable correlations with the occurrence of damage or the magnitude of the damage.

But even though the calculation of characteristic frequencies did not yield the best (though still good) results in this case, this method should not be disregarded in general. In this study, testing was conducted under laboratory conditions, and care was taken at all times to ensure that only one type of damage was present (tooth flank pitting). In real-world applications, damage can also occur on other components, such as bearings. Data-driven methods such as the calculation of

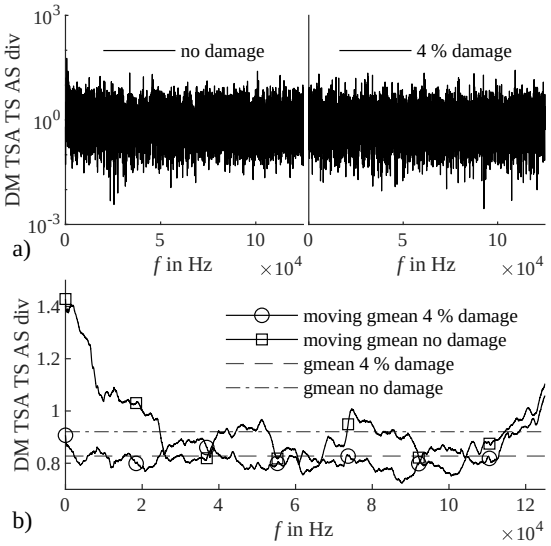

Figure 20. a) N-dimensional CI quotient of TS AS. b) Resulting CI TS AS geometric mean div. plotted as a moving RMS with a filter length of 10 % of the data points. Test run v12, gearbox 2.

difference spectra cannot distinguish between these types of damage. In contrast, the characteristic frequencies can be unambiguously attributed to the gearing.

To make a final assessment of the results, the levels according to Rytter (Rytter, 1993) are used again: **Damage detection** is possible at an early damage state using the presented methods. **Classifying the extent of the damage** remains difficult. The data points of the measurements with damage are clustered closely together, making it hard to distinguish between them.

New insights have also emerged when compared to the authors' previous work. In the straight-toothed spur gear transmissions examined in (Häderle et al., 2024; Merkle, Binanzer, Häderle, & Dazer, 2024), damage detection was still possible using characteristic frequencies. In particular, the physics-based approach involving the analysis of harmonics of the GMF still provided good results here. In planetary gears, however, this is much more difficult, and a data-driven approach must be adopted.

A key finding of this paper is that the correct selection of a CI is a decisive factor in the success of damage detection and, consequently, in the applicability of PHM strategies in gear drives. In the state of the art described in the literature, it has so far only been possible to detect very large damages ($\gg 4\%$). However, this study has shown that very early detection starting at 0.5% to 1% is already possible if the correct evaluation method is selected.

In general, it is important to note that a statistically significant effect does not necessarily mean that damage can be detected in practical applications. Although the two groups, y_{damage} and y_{healthy}, are statistically distinguishable, they also overlap to a large extent. Therefore, based on individual measurements from this group, it is not possible to assign them with absolute certainty. The recommendation for practical application must therefore be cyclical data collection at regular intervals in order to detect a shift in the group averages away from the healthy state in a early damage state. Only in this way an early damage detection in planetary gearboxes is possible.

6. SUMMARY AND CONCLUSION

While the state of the art has focused primarily on the detection of severe damage sizes, this research work focuses on the detection of damage in a very early damage state. By combining various methods of signal processing in the time and frequency domains, 130 resulting CIs were identified, and their ability to detect and classify damage was evaluated. Five resulting CIs have proven to be particularly suitable. These are examined in detail. A further 10 resulting CIs also show significant potential. This enables, in particular, very early detection. This represents a significant improvement over the state of the art on pitting damage detection. The compared basic CIs are not suitable for such early damage detection.

The findings obtained are, of course, initially valid for the gear system tested here. However, since the test gear systems are helical planetary gears, and thus represent the most kinematically complex gear design, the results are expected to be readily transferable to other gear types. Not every resulting CI exhibits the same sensitivity on every gearbox. Different CIs perform differently well on different gearboxes. As a recommendation for action, the simultaneous application of multiple parameters can be used to solve this challenge. Since no parameter in the presented study showed a pronounced tendency toward false positive values (as long as the calculation is performed within its own operating condition), multiple CIs can be monitored simultaneously without issues. If a change occurs in one of the monitored CIs, the occurrence of damage can be assumed.

In conclusion, it can be stated that traditional methods based on physical and kinematic principles (such as the analysis of characteristic frequencies) reach their limits when it comes to detecting very small damages. These methods are therefore unsuitable. Data-driven methods, such as the generation of difference spectra without preselecting individual frequencies, have provided significantly better results in this context.

This phenomenon can be further investigated in future work. Currently, no restriction on the frequency range has been applied. Following additional parameter studies, a meaningful restriction could be introduced here to enable the practical application of the method.

REFERENCES

Bechhoefer, E., & Kingsley, M. (2009). A Review of Time Synchronous Average Algorithms. In *Proceedings of the Annual Conference of the PHM Society 2009* (Vol. 1).

Bogert, B. P., Healy, M. J. R., & Tukey, J. W. (1963). The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum and Saphe Cracking. In *Proceedings of the Symposium on Time Series Analysis (M. Rosenblatt, Ed) Chapter 15* (pp. 209–143). New York, NY: Wiley.

Chaari, F., Fakhfakh, T., & Haddar, M. (2006, April). Dynamic analysis of a planetary gear failure caused by tooth pitting and cracking. *Journal of Failure Analysis and Prevention*, *6*(2), 73–78. doi: 10.1361/154770206X99343

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (Second Edition ed.). New York, NY: Routledge.

Farrar, C. R., & Doebling, S. W. (1999). Damage Detection and Evaluation II. In J. M. M. Silva & N. M. M. Maia (Eds.), *Modal Analysis and Testing* (pp. 345–378). Dordrecht: Springer Netherlands. doi: 10.1007/978-94-011-4503-9_17

Feng, Z., & Zuo, M. J. (2012, October). Vibration signal models for fault diagnosis of planetary gearboxes. *Journal of Sound and Vibration*, *331*(22), 4919–4939. doi: 10.1016/j.jsv.2012.05.039

FVA Forschungvereinigung Antriebstechnik e.V., & Binanzer, L. (2026). *FVA Heft 1698, FVA1036I, Sensorkonzept und Datenverarbeitung mittels Künstlicher Intelligenz zur Frühsterkennung von Schäden und deren Lokalisation in Getriebeanwendungen* (Forschungsvereinigung Antriebstechnik e.V., Ed.). Frankfurt.

Gretzinger, Y., Kroner, A., Henß, M., Dazer, M., & Bertsche, B. (2021, February). Extended Evaluation of Pitting Degradation Tests to Increase the Remaining Useful Life of Gear Wheels. *IOP Conference Series: Materials Science and Engineering*, *1097*(1), 012006. doi: 10.1088/1757-899X/1097/1/012006

Grzeszkowski, M., Gühmann, C., Scholzen, P., Löpenhaus, C., Nowoisky, S., & Kappmeyer, G. (2019). Experimental Study on the Pitting Detection Capabilities for Spur Gears Using Acoustic Emission and Vibration Analysis Methods. *GEARTECHNOLOGY*(March/April), 48–57.

Häderle, P., Merkle, L., & Dazer, M. (2024, December). Vibration analysis for early pitting detection during operation. *Forschung im Ingenieurwesen*, *88*(1), 15. doi: 10.1007/s10010-024-00743-5

Hou, J., Yang, S., Li, Q., & Liu, Y. (2022, January). Effect of a Novel Tooth Pitting Model on Mesh Stiffness and Vibration Response of Spur Gears. *Mathematics*, *10*(3), 471. doi: 10.3390/math10030471

Hu, C., Smith, W. A., Randall, R. B., & Peng, Z. (2016, August). Development of a gear vibration indicator and its application in gear wear monitoring. *Mechanical Systems and Signal Processing*, *76–77*, 319–336. doi: 10.1016/j.ymssp.2016.01.018

Inalpolat, M., & Kahraman, A. (2009, June). A theoretical and experimental investigation of modulation sidebands of planetary gear sets. *Journal of Sound and Vibration*, *323*(3-5), 677–696. doi: 10.1016/j.jsv.2009.01.004

ISO International Organization for Standardization. (2004). *ISO 18431-2, Mechanical vibration and shock — Signal processing — Part 2: Time domain windows for Fourier Transform analysis (ISO 18431-2:2004(E))*.

ISO International Organization for Standardization. (2007). *ISO 18431-4, Mechanical vibration and shock — Signal processing — Part 4: Shock-response spectrum analysis (ISO 18431-4:2007(E))*.

ISO International Organization for Standardization. (2014). *ISO 18431-3, Mechanical vibration and shock — Signal processing — Part 3: Methods of time-frequency analysis (ISO 18431-3:2014(E))*.

ISO International Organization for Standardization (Ed.). (2016). *ISO 6336-5, Calculation of load capacity of spur and helical gears — Part 5: Strength and quality of materials (ISO 6336-5:2016(E))*.

ISO International Organization for Standardization. (2020). *ISO 20816-9, Mechanical vibration — Measurement and evaluation of machine vibration — Part 9: Gear units (ISO 20816-9:2020(E))*.

Korka, Z. I., Bara, A., Clavac, B., & Filip, L. (2017). Gear Pitting Assessment Using Vibration Signal Analysis. *RJAV*, *vol XIV*(issue 1/2017), 44–49.

Kundu, P., Darpe, A. K., & Kulkarni, M. S. (2019, August). A correlation coefficient based vibration indicator for detecting natural pitting progression in spur gears. *Mechanical Systems and Signal Processing*, *129*, 741–763. doi: 10.1016/j.ymssp.2019.04.058

Kundu, P., Darpe, A. K., & Kulkarni, M. S. (2020, May). An ensemble decision tree methodology for remaining useful life prediction of spur gears under natural pitting progression. *Structural Health Monitoring*, *19*(3), 854–872. doi: 10.1177/1475921719865718

Kundu, P., Darpe, A. K., & Kulkarni, M. S. (2021, September). A review on diagnostic and prognostic approaches for gears. *Structural Health Monitoring*, *20*(5), 2853–2893. doi: 10.1177/1475921720972926

Kuzio, D., Zimroz, R., & Wyłomańska, A. (2023, August). Identification of fault frequency variation in the envelope spectrum in the vibration-based local damage detection in possible changing load/speed conditions. *Measurement*, *218*, 113148. doi: 10.1016/

j.measurement.2023.113148

Lebold, M., McClintic, K., Campbell, R., Byington, C., & Maynard, K. (2000). Review of Vibration Analysis Methods for Gearbox Diagnostics and Prognostics. In *Proceedings of the 54th Meeting of the Society for Machinery Failure Prevention Technology* (p. p. 623-634). Virginia Beach, VA: Applied Research Laboratory, The Pennsylvania State University.

Li, R., Xiong, X., Ma, J., & Zou, M. (2023, December). Effects of Macro-Pitting Fault on Dynamic Characteristics of Planetary Gear Train Considering Surface Roughness. *Actuators*, *13*(1), 1. doi: 10.3390/act13010001

Merkle, L., Binanzer, L., Häderle, P., & Dazer, M. (2024, November). Sensor concepts for gear damage detection in wind power drives. *Journal of Physics*, *2875*, 012022. doi: 10.1088/1742-6596/2875/1/012022

Merkle, Lukas, Binanzer, Lisa, & Dazer, Martin. (2025, September). Gear Damage Detection in Planetary Gearboxes. In *International Conference on Gears 2025*. Garching / Munich: VDI Verlag GmbH. doi: 10.51202/9783181024607

Norton, M. P. (2003). *Fundamentals of Noise and Vibration Analysis for Engineers* (2nd ed ed.). Cambridge, U.K New York: Cambridge University Press.

Ozturk, H., Yesilyurt, I., & Sabuncu, M. (2010, June). Detection and Advancement Monitoring of Distributed Pitting Failure in Gears. *Journal of Nondestructive Evaluation*, *29*(2), 63–73. doi: 10.1007/s10921-010-0066-4

Qu, Y., Zhang, Y., He, M., He, D., Jiao, C., & Zhou, Z. (2019, October). Gear pitting fault diagnosis using disentangled features from unsupervised deep learning. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, *233*(5), 719–730. doi: 10.1177/1748006X18822447

Randall, R. B. (2011). *Vibration-based Condition Monitoring*. John Wiley & Sons, Ltd.

Rytter, A. (1993). *Vibrational Based Inspection of Civil Engineering Structures* (Unpublished doctoral dissertation). Dept. of Building Technology and Structural Engineering, Aalborg University, Aalborg.

Sait, A. S., & Sharaf-Eldeen, Y. I. (2011). A Review of Gearbox Condition Monitoring Based on vibration Analysis Techniques Diagnostics and Prognostics. In T. Proulx (Ed.), *Rotating Machinery, Structural Health Monitoring, Shock and Vibration, Volume 5* (pp. 307–324). New York, NY: Springer New York. doi: 10.1007/978-1-4419-9428-8_25

Sánchez, R.-V., Lucero, P., Vásquez, R. E., Cerrada, M., & Cabrera, D. (2018). A comparative feature analysis for gear pitting level classification by using acoustic emission, vibration and current signals. *IFAC-PapersOnLine*, *51*(24), 346–352. doi: 10.1016/j.ifacol.2018.09.600

Sarvestani, E. S., Rezaeizadeh, M., Jomehzadeh, E., & Bigani, A. (2020, June). Early Detection of Industrial-Scale Gear Tooth Surface Pitting Using Vibration Analysis. *Journal of Failure Analysis and Prevention*, *20*(3), 768–788. doi: 10.1007/s11668-020-00874-1

Sendlbeck, S., Fromberger, M., Otto, M., & Stahl, K. (2021). Vibration-based Gear Condition Monitoring using an Improved Section-specific Approach without the need of Historic Reference Data.

Teng, W., Wang, F., Zhang, K., Liu, Y., & Ding, X. (2014, January). Pitting Fault Detection of a Wind Turbine Gearbox Using Empirical Mode Decomposition. *Strojniški vestnik – Journal of Mechanical Engineering*, *60*(1), 12–20. doi: 10.5545/sv-jme.2013.1295

Tevetoglu, T., Henss, M., Gretzinger, Y., & Bertsche, B. (2021, November). Learning an Optimal Operational Strategy for Service Life Extension of Gear Wheels with Double Deep Q Networks. *Annual Conference of the PHM Society*, *13*(1). doi: 10.36001/phmconf.2021.v13i1.2978

The MathWorks, Inc. (2026). pwelch - Welch's power spectral density estimate. *MATLAB Help Center*. https://de.mathworks.com/help/signal/ref/pwelch.html.

Vishwakarma, M., Purohit, R., Harshlata, V., & Rajput, P. (2017). Vibration Analysis & Condition Monitoring for Rotating Machines: A Review. *Materials Today: Proceedings*, *4*(2), 2659–2664. doi: 10.1016/j.matpr.2017.02.140

Welch, P. (1967, June). The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms. *IEEE Transactions on Audio and Electroacoustics*, *15*(2), 70–73. doi: 10.1109/TAU.1967.1161901

Yakeu Happi, K. H., Kouejou, B. X. T., & Anyika Alugongo, A. (2023, December). Experimental study and comparative analysis of pitting fault in spur gear system. *Journal of Vibroengineering*, *25*(8), 1480–1501. doi: 10.21595/jve.2023.23335

Zhe, C., Niaoqing, H., Fengshou, G., & Guojun, Q. (2011, September). Pitting Damage Levels Estimation for Planetary Gear Sets Based on Model Simulation and Grey Relational Analysis. *Transactions of the Canadian Society for Mechanical Engineering*, *35*(3), 403–417. doi: 10.1139/tcsme-2011-0023

Zhu, R., Maele, D. V., Poletto, J. C., Baets, P. D., & Gryllias, K. (2024). Development and evaluation of vibration indicators for gear pitting detection. In *International Conference on Noise and Vibration Engineering (ISMA 2024) and International Conference on Uncertainty in Structural Dynamics (USD 2024)* (pp. 508–521). Leuven, Belgium: Katholieke Universiteit Leuven, Department of Mechanical Engineering.

APPENDIX

Table 3. Result of t-test. All statistically significant CIs with effect size $d > 0.5$

No.	CI	Description	Effect size d
a)	DM TSA TS AS gmean div	The geometric mean quotient of the CI amplitude spectrum of the demodulated TSA time signal without DC bias	0.969
b)	DM TS AS mean dif	The mean difference of the CI amplitude spectrum of the demodulated time signal without DC bias	0.946
c)	DM TS AS gmean div	The geometric mean quotient of the CI amplitude spectrum of the demodulated time signal without DC bias	0.944
d)	TS AS mean dif	The mean difference of the CI amplitude spectrum of the time signal	0.882
e)	TS AS gmean div	The geometric mean quotient of the CI amplitude spectrum of the time signal	0.817
	DM TS AS CFm mean dif	The mean difference of the CI characteristic frequencies in amplitude spectrum of the demodulated time signal without DC bias	0.790
	DM TS AS CFm gmean div	The geometric mean quotient of the CI characteristic frequencies in amplitude spectrum of the demodulated time signal without DC bias	0.758
	TSA TS AS gmean div	The geometric mean quotient of the CI amplitude spectrum of the TSA time signal	0.689
	DM TSA TS AS mean dif	The mean difference of the CI amplitude spectrum of the demodulated TSA time signal without DC bias	0.643
	TS AS CFm mean dif	The mean difference of the CI characteristic frequencies in amplitude spectrum of the time signal	0.630
	DM TSA TS AS RMS div	The RMS quotient of the CI amplitude spectrum of the demodulated TSA time signal without DC bias	0.628
	DM TS AS CFm RMS div	The RMS quotient of the CI characteristic frequencies in amplitude spectrum of the demodulated time signal without DC bias	0.588
	DM TS AS RMS div	The RMS quotient of the CI amplitude spectrum of the demodulated time signal without DC bias	0.578
	TS AS CFm gmean div	The geometric mean quotient of the CI characteristic frequencies in amplitude spectrum of the time signal	0.535
	DM TS AS cosSim	The cos. similarity of the CI amplitude spectrum of the demodulated time signal without DC bias	0.516